

Policy Evaluation with Variance Related Risk Criteria in Markov Decision Processes

Aviv Tamar
Dotan Di Castro
Shie Mannor

AVIVT@TX.TECHNION.AC.IL
DOT@TX.TECHNION.AC.IL
SHIE@EE.TECHNION.AC.IL

Department of Electrical Engineering, The Technion - Israel Institute of Technology, Haifa, Israel 32000

Abstract

In this paper we extend temporal difference policy evaluation algorithms to performance criteria that include the variance of the cumulative reward. Such criteria are useful for risk management, and are important in domains such as finance and process control. We propose both TD(0) and LSTD(λ) variants with linear function approximation, prove their convergence, and demonstrate their utility in a 4-dimensional continuous state space problem.

1. Introduction

In both Reinforcement Learning (RL; Bertsekas & Tsitsiklis, 1996) and planning in Markov Decision Processes (MDPs; Puterman, 1994), the typical objective is to maximize the cumulative (possibly discounted) expected reward, denoted by J . In many applications, however, the decision maker is also interested in minimizing some form of *risk* of the policy. By risk, we mean reward criteria that take into account not only the expected reward, but also some additional statistics of the total reward such as its variance, its Value at Risk, etc. (Luenberger, 1998).

In this work we focus on risk measures that involve the *variance of the cumulative reward*, denoted by V . Typical performance criteria that fall under this definition include

- (a) Maximize J s.t. $V \leq c$
- (b) Minimize V s.t. $J \geq c$
- (c) Maximize the Sharpe Ratio: J/\sqrt{V}

- (d) Maximize $J - c\sqrt{V}$

The rationale behind our choice of risk measure is that these performance criteria, such as the Sharpe Ratio (Sharpe, 1966) mentioned above, are being used in practice. Moreover, it seems that human decision makers understand how to use variance well, in comparison to exponential utility functions (Howard & Matheson, 1972), which require determining a non-intuitive exponent coefficient.

A fundamental concept in RL is the the value function - the expected reward to go from a given state. Estimates of the value function drive most RL algorithms, and efficient methods for obtaining these estimates have been a prominent area of research. In particular, Temporal Difference (TD; (Sutton & Barto, 1998)) based methods have been found suitable for problems where the state space is large, requiring some sort of function approximation. TD methods enjoy theoretical guarantees (Bertsekas, 2012; Lazaric et al., 2010) and empirical success (Tesauro, 1995), and are considered the state of the art in policy evaluation.

In this work we present a TD framework for estimating the *variance of the reward to go*. Our approach is based on the following key observation: the second moment of the reward to go, denoted by M , together with the value function J , obey a linear equation - similar to the Bellman equation that drives regular TD algorithms. By extending TD methods to jointly estimate J and M , we obtain a solution for estimating the variance, using the relation $V = M - J^2$.

We propose both a variant of Least Squares Temporal Difference (LSTD) (Boyan, 2002) and of TD(0) (Sutton & Barto, 1998) for jointly estimating J and M with a linear function approximation. For these algorithms, we provide convergence guarantees and error bounds. In addition, we introduce a novel approach for enforcing the approximate variance to be positive, through a constrained TD equation.

Finally, an empirical evaluation on a challenging continuous maze domain highlights both the usefulness of our approach, and the importance of the variance function in understanding the risk of a policy.

This paper is organized as follows. In Section 2 we present our formal RL setup. In Section 3 we derive the fundamental equations for jointly approximating J and M , and discuss their properties. A solution to these equations may be obtained by simulation, through the use of TD algorithms, as presented in Section 4. In Section 5 we further extend the LSTD framework by forcing the approximated variance to be positive. Section 6 presents an empirical evaluation, and Section 7 concludes, and discusses future directions.

2. Framework and Background

We consider a Stochastic Shortest Path (SSP) problem¹ (Bertsekas, 2012), where the environment is modeled by an MDP in discrete time with a finite state set $X \triangleq \{1, \dots, n\}$ and a terminal state x^* . A fixed policy π determines, for each $x \in X$, a stochastic transition to a subsequent state $y \in \{X \cup x^*\}$ with probability $P(y|x)$. We consider a deterministic and bounded reward function $r : X \rightarrow \mathbb{R}$. We denote by x_k the state at time k , where $k = 0, 1, 2, \dots$

A policy is said to be *proper* (Bertsekas, 2012) if there is a positive probability that the terminal state x^* will be reached after at most n transitions, from any initial state. In this paper we make the following assumption

Assumption 1. *The policy π is proper.*

Let $\tau \triangleq \min\{k > 0 | x_k = x^*\}$ denote the first visit time to the terminal state, and let the random variable B denote the accumulated reward along the trajectory until that time²

$$B \triangleq \sum_{k=0}^{\tau-1} r(x_k).$$

In this work, we are interested in the mean-variance tradeoff in B , represented by the *value function*

$$J(x) \triangleq \mathbb{E}[B|x_0 = x], \quad x \in X,$$

and the *variance of the reward to go*

$$V(x) \triangleq \text{Var}[B|x_0 = x], \quad x \in X.$$

¹This is also known as an episodic setup.

²We do not define the reward at the terminal state as it is not relevant to our performance criteria. However, the customary zero terminal reward may be assumed throughout the paper.

We will find it convenient to define also the *second moment of the reward to go*

$$M(x) \triangleq \mathbb{E}[B^2|x_0 = x], \quad x \in X.$$

Our goal is to estimate $J(x)$ and $V(x)$ from trajectories obtained by simulating the MDP with policy π .

3. Approximation of the Variance of the Reward To Go

In this section we derive a projected equation method for approximating $J(x)$ and $M(x)$ using linear function approximation. The estimation of $V(x)$ will then follow from the relation $V(x) = M(x) - J(x)^2$.

Our starting point is a system of equations for $J(x)$ and $M(x)$, first derived by Sobel (1982) for a discounted infinite horizon case, and extended here to the SSP case. Note that the equation for J is the well known Bellman equation for a fixed policy, and independent of the equation for M .

Proposition 2. *The following equations hold for $x \in X$*

$$\begin{aligned} J(x) &= r(x) + \sum_{y \in X} P(y|x)J(y), \\ M(x) &= r(x)^2 + 2r(x) \sum_{y \in X} P(y|x)J(y) + \sum_{y \in X} P(y|x)M(y). \end{aligned} \quad (1)$$

Furthermore, under Assumption 1 a unique solution to (1) exists.

The proof is straightforward, and given in Appendix A.

At this point the reader may wonder why an equation for V is not presented. While such an equation may be derived, as was done in (Tamar et al., 2012), it is not linear. The linearity of (1) is the key to our approach. As we show in the next subsection, the solution to (1) may be expressed as the fixed point of a linear mapping in the joint space of J and M . We will then show that a projection of this mapping onto a linear feature space is contracting, thus allowing us to use existing TD theory to derive estimation algorithms for J and M .

3.1. A Projected Fixed Point Equation on the Joint Space of J and M

For the sequel we introduce the following vector notations. We denote by $P \in \mathbb{R}^{n \times n}$ and $r \in \mathbb{R}^n$ the SSP transition matrix and reward vector, i.e.,

$P_{x,y} = P(y|x)$ and $r_x = r(x)$, where $x, y \in X$. Also, we define $R \triangleq \text{diag}(r)$.

For a vector $z \in \mathbb{R}^{2n}$ we let $z_J \in \mathbb{R}^n$ and $z_M \in \mathbb{R}^n$ denote its leading and ending n components, respectively. Thus, such a vector belongs to the joint space of J and M .

We define the mapping $T : \mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$ by

$$\begin{aligned} [Tz]_J &= r + Pz_J, \\ [Tz]_M &= Rr + 2RPz_J + Pz_M. \end{aligned}$$

It may easily be verified that a fixed point of T is a solution to (1), and by Proposition 2 such a fixed point exists and is unique.

When the state space X is large, a direct solution of (1) is not feasible, even if P may be accurately obtained. A popular approach in this case is to approximate $J(x)$ by restricting it to a lower dimensional subspace, and use simulation based TD algorithms to adjust the approximation parameters (Bertsekas, 2012). In this paper we extend this approach to the approximation of $M(x)$ as well.

We consider a linear approximation architecture of the form

$$\begin{aligned} \tilde{J}(x) &= \phi_J(x)^T w_J, \\ \tilde{M}(x) &= \phi_M(x)^T w_M, \end{aligned} \quad (2)$$

where $w_J \in \mathbb{R}^{l_J}$ and $w_M \in \mathbb{R}^{l_M}$ are the approximation parameter vectors, $\phi_J(x) \in \mathbb{R}^{l_J}$ and $\phi_M(x) \in \mathbb{R}^{l_M}$ are state dependent features, and $(\cdot)^T$ denotes the transpose of a vector. The low dimensional subspaces are therefore

$$\begin{aligned} S_J &= \{\Phi_J w | w \in \mathbb{R}^{s_J}\}, \\ S_M &= \{\Phi_M w | w \in \mathbb{R}^{s_M}\}, \end{aligned}$$

where Φ_J and Φ_M are matrices whose rows are $\phi_J(x)^T$ and $\phi_M(x)^T$, respectively. We make the following standard independence assumption on the features

Assumption 3. *The matrix Φ_J has rank l_J and the matrix Φ_M has rank l_M .*

As outlined earlier, our goal is to estimate w_J and w_M from simulated trajectories of the MDP. Thus, it is constructive to consider projections onto S_J and S_M with respect to a norm that is weighted according to the state occupancy in these trajectories.

For a trajectory $x_0, \dots, x_{\tau-1}$, where x_0 is drawn from a fixed distribution $\zeta_0(x)$, and the states evolve according to the MDP with policy π , define the state occupancy probabilities

$$q_t(x) = P(x_t = x), \quad x \in X, \quad t = 0, 1, \dots$$

and let

$$\begin{aligned} q(x) &= \sum_{t=0}^{\infty} q_t(x), \quad x \in X \\ Q &\triangleq \text{diag}(q). \end{aligned}$$

We make the following assumption on the policy π and initial distribution ζ_0

Assumption 4. *Each state has a positive probability of being visited, namely, $q(x) > 0$ for all $x \in X$.*

For vectors in \mathbb{R}^n , we introduce the weighted Euclidean norm

$$\|y\|_q = \sqrt{\sum_{i=1}^n q(i) (y(i))^2}, \quad y \in \mathbb{R}^n,$$

and we denote by Π_J and Π_M the projections from \mathbb{R}^n onto the subspaces S_J and S_M , respectively, with respect to this norm. For $z \in \mathbb{R}^{2n}$ we denote by Π the projection of z_J onto S_J and z_M onto S_M , namely³

$$\Pi = \begin{pmatrix} \Pi_J & 0 \\ 0 & \Pi_M \end{pmatrix}. \quad (3)$$

We are now ready to fully describe our approximation scheme. We consider the *projected* fixed point equation

$$z = \Pi T z, \quad (4)$$

and, letting z^* denote its solution, propose the approximate value function $\tilde{J} = z_J^* \in S_J$ and second moment function $\tilde{M} = z_M^* \in S_M$.

We proceed to derive some properties of the projected fixed point equation (4). We begin by stating a well known result regarding the contraction properties of the *projected Bellman operator* $\Pi_J T_J$, where $T_J y = r + Py$. A proof can be found at (Bertsekas, 2012), proposition 7.1.1.

Lemma 5. *Let Assumptions 1, 3, and 4 hold. Then, there exists some norm $\|\cdot\|_J$ and some $\beta_J < 1$ such that*

$$\|\Pi_J P y\|_J \leq \beta_J \|y\|_J, \quad \forall y \in \mathbb{R}^n.$$

Similarly, there exists some norm $\|\cdot\|_M$ and some $\beta_M < 1$ such that

$$\|\Pi_M P y\|_M \leq \beta_M \|y\|_M, \quad \forall y \in \mathbb{R}^n.$$

Next, we define a weighted norm on \mathbb{R}^{2n}

³The projection operators Π_J and Π_M are linear, and may be written explicitly as $\Pi_J = \Phi_J(\Phi_J^T Q \Phi_J)^{-1} \Phi_J^T Q$, and similarly for Π_M .

Definition 6. For a vector $z \in \mathbb{R}^{2n}$ and a scalar $0 < \alpha < 1$, the α -weighted norm is

$$\|z\|_\alpha = \alpha\|z_J\|_J + (1 - \alpha)\|z_M\|_M, \quad (5)$$

where the norms $\|\cdot\|_J$ and $\|\cdot\|_M$ are defined in Lemma 5.

Our main result of this section is given in the following lemma, where we show that the projected operator ΠT is a contraction with respect to the α -weighted norm.

Lemma 7. Let Assumptions 1, 3, and 4 hold. Then, there exists some $0 < \alpha < 1$ and some $\beta < 1$ such that ΠT is a β -contraction with respect to the α -weighted norm, i.e.,

$$\|\Pi T z\|_\alpha \leq \beta\|z\|_\alpha, \quad \forall z \in \mathbb{R}^{2n}.$$

Proof. Let \mathcal{P} denote the following matrix in $\mathbb{R}^{2n \times 2n}$

$$\mathcal{P} = \begin{pmatrix} P & 0 \\ 2RP & P \end{pmatrix},$$

and let $z \in \mathbb{R}^{2n}$. We need to show that

$$\|\Pi \mathcal{P} z\|_\alpha \leq \beta\|z\|_\alpha.$$

From (3) we have

$$\Pi \mathcal{P} = \begin{pmatrix} \Pi_J P & 0 \\ 2\Pi_M R P & \Pi_M P \end{pmatrix}.$$

Therefore, we have

$$\begin{aligned} \|\Pi \mathcal{P} z\|_\alpha &= \alpha\|\Pi_J P z_J\|_J \\ &\quad + (1 - \alpha)\|2\Pi_M R P z_J + \Pi_M P z_M\|_M \\ &\leq \alpha\|\Pi_J P z_J\|_J \\ &\quad + (1 - \alpha)\|\Pi_M P z_M\|_M \\ &\quad + (1 - \alpha)\|2\Pi_M R P z_J\|_M \\ &\leq \alpha\beta_J\|z_J\|_J \\ &\quad + (1 - \alpha)\beta_M\|z_M\|_M \\ &\quad + (1 - \alpha)\|2\Pi_M R P z_J\|_M, \end{aligned} \quad (6)$$

where the equality is by definition of the α weighted norm (5), the first inequality is from the triangle inequality, and the second inequality is by Lemma 5. Now, we claim that there exists some finite C such that

$$\|2\Pi_M R P y\|_M \leq C\|y\|_J, \quad \forall y \in \mathbb{R}^n. \quad (7)$$

To see this, note that since \mathbb{R}^n is a finite dimensional real vector space, all vector norms are equivalent (Horn & Johnson, 1985) therefore there exist finite C_1 and C_2 such that for all $y \in \mathbb{R}^n$

$$C_1\|2\Pi_M R P y\|_2 \leq \|2\Pi_M R P y\|_M \leq C_2\|2\Pi_M R P y\|_2,$$

where $\|\cdot\|_2$ denotes the Euclidean norm. Let λ denote the spectral norm of the matrix $2\Pi_M R P$, which is finite since all the matrix elements are finite. We have

$$\|2\Pi_M R P y\|_2 \leq \lambda\|y\|_2, \quad \forall y \in \mathbb{R}^n.$$

Using again the fact that all vector norms are equivalent, there exists a finite C_3 such that

$$\|y\|_2 \leq C_3\|y\|_J, \quad \forall y \in \mathbb{R}^n.$$

Setting $C = C_2\lambda C_3$ we get the desired bound. Let $\beta = \max\{\beta_J, \beta_M\} < 1$, and choose $\epsilon > 0$ such that

$$\tilde{\beta} + \epsilon < 1.$$

Now, choose α such that

$$\alpha = \frac{C}{\epsilon + C}.$$

We have that

$$(1 - \alpha)C = \alpha\epsilon,$$

and plugging in (7)

$$(1 - \alpha)\|2\Pi_M R P y\|_M \leq \alpha\epsilon\|y\|_J.$$

Plugging in (6) we have

$$\begin{aligned} &\alpha\beta_J\|z_J\|_J + (1 - \alpha)\beta_M\|z_M\|_M + (1 - \alpha)\|2\Pi_M R P z_J\|_M \\ &\leq \alpha\beta_J\|z_J\|_J + (1 - \alpha)\beta_M\|z_M\|_M + \alpha\epsilon\|z_J\|_J \\ &\leq (\tilde{\beta} + \epsilon)(\alpha\|z_J\|_J + (1 - \alpha)\|z_M\|_M) \end{aligned}$$

and therefore

$$\|\Pi \mathcal{P} z\|_\alpha \leq (\tilde{\beta} + \epsilon)\|z\|_\alpha$$

Finally, choose $\beta = \tilde{\beta} + \epsilon$. \square

Lemma 7 guarantees that the projected operator ΠT has a unique fixed point. Let us denote this fixed point by z^* , and let w_J^*, w_M^* denote the corresponding weights, which are unique due to Assumption 3

$$\begin{aligned} \Pi T z^* &= z^*, \\ z_J^* &= \Phi_J w_J^*, \\ z_M^* &= \Phi_M w_M^*. \end{aligned} \quad (8)$$

In the next lemma we provide a bound on the approximation error. The proof is in Appendix B.

Lemma 8. Let Assumptions 1, 3, and 4 hold. Denote by $z_{true} \in \mathbb{R}^{2n}$ the true value and second moment functions, i.e., z_{true} satisfies $z_{true} = T z_{true}$. Then,

$$\|z_{true} - z^*\|_\alpha \leq \frac{1}{1 - \beta}\|z_{true} - \Pi z_{true}\|_\alpha,$$

with α and β defined in Lemma 7.

4. Simulation Based Estimation Algorithms

We now use the theoretical results of the previous subsection to derive simulation based algorithms for jointly estimating the value function and second moment. The projected equation (8) is linear, and can be written in matrix form as follows. First let us write the equation explicitly as

$$\begin{aligned} \Pi_J (r + P\Phi_J w_J^*) &= \Phi_J w_J^*, \\ \Pi_M (Rr + 2RP\Phi_J w_J^* + P\Phi_M w_M^*) &= \Phi_M w_M^*. \end{aligned} \quad (9)$$

Projecting a vector y onto Φw satisfies the following orthogonality condition

$$\Phi^T Q(y - \Phi w) = 0,$$

therefore we have

$$\begin{aligned} \Phi_J^T Q(\Phi_J w_J^* - (r + P\Phi_J w_J^*)) &= 0, \\ \Phi_M^T Q(\Phi_M w_M^* - (Rr + 2RP\Phi_J w_J^* + P\Phi_M w_M^*)) &= 0, \end{aligned}$$

which can be written as

$$\begin{aligned} Aw_J^* &= b, \\ Cw_M^* &= d, \end{aligned} \quad (10)$$

with

$$\begin{aligned} A &= \Phi_J^T Q(I - P)\Phi_J, \quad b = \Phi_J^T Qr, \\ C &= \Phi_M^T Q(I - P)\Phi_M, \quad d = \Phi_M^T QR(r + 2P\Phi_J A^{-1}b), \end{aligned} \quad (11)$$

and the matrices A and C are invertible since Lemma 7 guarantees a unique solution to (8) and Assumption 3 guarantees the unique weights of its projection.

4.1. A Least Squares TD Algorithm

Our first simulation based algorithm is an extension of the Least Squares Temporal Difference (LSTD) algorithm (Boyan, 2002). We simulate N trajectories of the MDP with the policy π and initial state distribution ζ_0 . Let $x_0^k, x_1^k, \dots, x_{\tau^k-1}^k$ and τ^k , where $k = 0, 1, \dots, N$, denote the state sequence and visit times to the terminal state within these trajectories, respectively. We now use these trajectories to form the

following estimates of the terms in (11)

$$\begin{aligned} A_N &= \mathbb{E}_N \left[\sum_{t=0}^{\tau-1} \phi_J(x_t)(\phi_J(x_t) - \phi_J(x_{t+1}))^T \right], \\ b_N &= \mathbb{E}_N \left[\sum_{t=0}^{\tau-1} \phi_J(x_t)r(x_t) \right], \\ C_N &= \mathbb{E}_N \left[\sum_{t=0}^{\tau-1} \phi_M(x_t)(\phi_M(x_t) - \phi_M(x_{t+1}))^T \right], \\ d_N &= \mathbb{E}_N \left[\sum_{t=0}^{\tau-1} \phi_M(x_t)r(x_t) (r(x_t) + 2\phi_J(x_{t+1})^T A_N^{-1}b_N) \right], \end{aligned} \quad (12)$$

where \mathbb{E}_N denotes an empirical average over trajectories, i.e., $\mathbb{E}_N[f(x, \tau)] = \frac{1}{N} \sum_{k=1}^N f(x^k, \tau^k)$. The LSTD approximation is given by

$$\begin{aligned} \hat{w}_J^* &= A_N^{-1}b_N, \\ \hat{w}_M^* &= C_N^{-1}d_N. \end{aligned}$$

The next theorem shows that the LSTD approximation converges.

Theorem 9. *Let Assumptions 1, 3, and 4 hold. Then $\hat{w}_J^* \rightarrow w_J^*$ and $\hat{w}_M^* \rightarrow w_M^*$ as $N \rightarrow \infty$ with probability 1.*

The proof involves a straightforward application of the law of large numbers and is described in Appendix C.

4.2. An online TD(0) Algorithm

Our second estimation algorithm is an extension of the well known TD(0) algorithm (Sutton & Barto, 1998). Again, we simulate trajectories of the MDP corresponding to the policy π and initial state distribution ζ_0 , and we iteratively update our estimates at every visit to the terminal state⁴. For some $0 \leq t < \tau^k$ and weights w_J, w_M , we introduce the TD terms

$$\begin{aligned} \delta_J^k(t, w_J, w_M) &= r(x_t^k) + (\phi_J(x_{t+1}^k)^T - \phi_J(x_t^k)^T) w_J, \\ \delta_M^k(t, w_J, w_M) &= r^2(x_t^k) + 2r(x_t^k)\phi_J(x_{t+1}^k)^T w_J \\ &\quad + (\phi_M(x_{t+1}^k)^T - \phi_M(x_t^k)^T) w_M. \end{aligned}$$

Note that δ_J^k is the standard TD error (Sutton & Barto, 1998). The TD(0) update is given by

$$\begin{aligned} \hat{w}_{J;k+1} &= \hat{w}_{J;k} + \xi_k \sum_{t=0}^{\tau^k-1} \phi_J(x_t) \delta_J^k(t, \hat{w}_{J;k}, \hat{w}_{M;k}), \\ \hat{w}_{M;k+1} &= \hat{w}_{M;k} + \xi_k \sum_{t=0}^{\tau^k-1} \phi_M(x_t) \delta_M^k(t, \hat{w}_{J;k}, \hat{w}_{M;k}), \end{aligned}$$

⁴An extension to an algorithm that updates at every state transition is also possible, but we do not pursue such here.

where $\{\xi_k\}$ are positive step sizes.

The next theorem shows that the TD(0) algorithm converges.

Theorem 10. *Let Assumptions 1, 3, and 4 hold, and let the step sizes satisfy*

$$\sum_{k=0}^{\infty} \xi_k = \infty, \quad \sum_{k=0}^{\infty} \xi_k^2 < \infty.$$

Then $\hat{w}_{J;k} \rightarrow w_J^$ and $\hat{w}_{M;k} \rightarrow w_M^*$ as $k \rightarrow \infty$ with probability 1.*

The proof, provided in Appendix D, is based on representing the TD(0) algorithm as a stochastic approximation and using contraction properties similar to the ones of the previous section to prove convergence.

4.3. Multistep Algorithms

A common method in value function approximation is to replace the single step mapping T_J with a multistep version of the form

$$T_J^{(\lambda)} = (1 - \lambda) \sum_{l=0}^{\infty} \lambda^l T_J^{l+1}$$

with $0 < \lambda < 1$. The projected equation (9) then becomes

$$\Pi_J T_J^{(\lambda)} (\Phi_J w_J^{*(\lambda)}) = \Phi_J w_J^{*(\lambda)}.$$

Similarly, we may write a multistep equation for M

$$\Pi_M T_M^{(\lambda)} (\Phi_M w_M^{*(\lambda)}) = \Phi_M w_M^{*(\lambda)}, \quad (13)$$

where

$$T_M^{(\lambda)} = (1 - \lambda) \sum_{l=0}^{\infty} \lambda^l T_{M^*}^{l+1},$$

and

$$T_{M^*}(y) = Rr + 2RP\Phi_J w_J^{*(\lambda)} + Py.$$

Note the difference between T_{M^*} and T_M defined earlier; We are no longer working on the joint space of J and M but instead we have an independent equation for approximating J , and its solution $w_J^{*(\lambda)}$ is part of equation (13) for approximating M . By Proposition 7.1.1. of (Bertsekas, 2012) both $\Pi_J T_J^{(\lambda)}$ and $\Pi_M T_M^{(\lambda)}$ are contractions with respect to the weighted norm $\|\cdot\|_q$, therefore both multistep projected equations admit a unique solution. In a similar manner to the single step version, the projected equations may be written in matrix form

$$\begin{aligned} A^{(\lambda)} w_J^{*(\lambda)} &= b^{(\lambda)}, \\ C^{(\lambda)} w_M^{*(\lambda)} &= d^{(\lambda)}, \end{aligned} \quad (14)$$

where

$$A^{(\lambda)} = \Phi_J^T Q (I - P^{(\lambda)}) \Phi_J, \quad b^{(\lambda)} = \Phi_J^T Q (I - \lambda P)^{-1} r,$$

$$C^{(\lambda)} = \Phi_M^T Q (I - P^{(\lambda)}) \Phi_M,$$

$$d^{(\lambda)} = \Phi_M^T Q (I - \lambda P)^{-1} R (r + 2P\Phi_J w_J^{*(\lambda)}),$$

and

$$P^{(\lambda)} = (1 - \lambda) \sum_{l=0}^{\infty} \lambda^l P^{l+1}.$$

Simulation based estimates $A_N^{(\lambda)}$ and $b_N^{(\lambda)}$ of the expressions above may be obtained by the use of eligibility traces, as described in (Bertsekas, 2012), and the LSTD(λ) approximation is then given by $\hat{w}_J^{*(\lambda)} = (A_N^{(\lambda)})^{-1} b_N^{(\lambda)}$. By substituting $w_J^{*(\lambda)}$ with $\hat{w}_J^{*(\lambda)}$ in the expression for $d^{(\lambda)}$, a similar procedure may be used to derive estimates $C_N^{(\lambda)}$ and $d_N^{(\lambda)}$, and to obtain the LSTD(λ) approximation $\hat{w}_M^{*(\lambda)} = (C_N^{(\lambda)})^{-1} d_N^{(\lambda)}$. Due to the similarity to the LSTD procedure in (12), the exact details are omitted.

5. Positive Variance as a Constraint in LSTD

The TD algorithms of the preceding section approximated J and M by the solution to the fixed point equation (8). While Lemma 8 provides us a bound on the approximation error of \tilde{J} and \tilde{M} measured in the α -weighted norm, it does not guarantee that the approximated variance \tilde{V} , given by $\tilde{M} - \tilde{J}^2$, is positive for all states. If we are estimating M as a means to infer V , it may be useful to include our prior knowledge that $V \geq 0$ in the estimation process. In this section we propose to enforce this knowledge as a constraint in the projected fixed point equation.

The multistep equation for the second moment weights (13) may be written with the projection operator as an explicit minimization

$$w_M^{*(\lambda)} = \arg \min_w \|\Phi_M w - (\tilde{r} + \tilde{\Phi} w_M^{*(\lambda)})\|_q,$$

with

$$\tilde{\Phi} = P^{(\lambda)} \Phi_M,$$

and

$$\tilde{r} = (I - \lambda P)^{-1} (Rr + 2RP\Phi_J w_J^{*(\lambda)}).$$

Requiring non negative variance in some state x may be written as a linear constraint in $w_M^{*(\lambda)}$

$$\phi_M(x)^T w_M^{*(\lambda)} - (\phi_J(x)^T w_J^{*(\lambda)})^2 \geq 0.$$

Let $\{x_1, \dots, x_l\}$ denote a set of states in which we demand that the variance be non negative. Let $H \in \mathbb{R}^{l \times l_M}$ denote a matrix with the features $-\phi_M^T(x_i)$ as its rows, and let $g \in \mathbb{R}^l$ denote a vector with elements $-(\phi_J(x_i)^T w_J^{*(\lambda)})^2$. We can write the variance-constrained projected equation for the second moment as

$$w_M^{vc} = \begin{cases} \arg \min_w & \|\Phi_M w - (\tilde{r} + \tilde{\Phi} w_M^{vc})\|_q \\ \text{s.t.} & Hw \leq g \end{cases} \quad (15)$$

The following assumption guarantees that the constraints in (15) admit a feasible solution.

Assumption 11. *There exists w such that $Hw < g$.*

Note that a simple way to satisfy Assumption 11 is to have some feature vector that is positive for all states. Equation (15) is a form of projected equation studied in (Bertsekas, 2011), the solution of which may be obtained by the following iterative procedure

$$w_{k+1} = \Pi_{\Xi, \hat{W}_M} [w_k - \gamma \Xi^{-1} (C^{(\lambda)} w_k - d^{(\lambda)})], \quad (16)$$

where Ξ is some positive definite matrix, and Π_{Ξ, \hat{W}_M} denotes a projection onto the convex set $\hat{W}_M = \{w | Hw \leq g\}$ with respect to the Ξ weighted Euclidean norm. The following lemma, which is based on a convergence result of (Bertsekas, 2011), guarantees that algorithm (16) converges.

Lemma 12. *Assume $\lambda > 0$. Then there exists $\bar{\gamma} > 0$ such that $\forall \gamma \in (0, \bar{\gamma})$ the algorithm (16) converges at a linear rate to w_M^{vc} .*

Proof. This is a direct application of the convergence result in (Bertsekas, 2011). The only nontrivial assumption that needs to be verified is that $T_M^{(\lambda)}$ is a contraction in the $\|\cdot\|_q$ norm (Proposition 1 in Bertsekas, 2011). For $\lambda > 0$ Proposition 7.1.1. of (Bertsekas, 2012) guarantees that $T_M^{(\lambda)}$ is indeed contracting in the $\|\cdot\|_q$ norm. \square

We illustrate the effect of the positive variance constraint in a simple example. Consider the Markov chain depicted in Figure 1, which consists of N states with reward -1 and a terminal state x^* with zero reward. The transitions from each state is either to a subsequent state (with probability p) or to a preceding state (with probability $1-p$), with the exception of the first state which transitions to itself instead. We chose to approximate J and M with polynomials of degree 1 and 2, respectively. For such a small problem the fixed point equation (14) may be solved exactly, yielding the approximation depicted in Figure 2 (dotted line), for $p = 0.7$, $N = 30$, and $\lambda = 0.95$. Note

that the variance is negative for the last two states. Using algorithm (16) we obtained a positive variance constrained approximation, which is depicted in figure 2 (dashed line). Note that the variance is now positive for all states (as was required by the constraints).

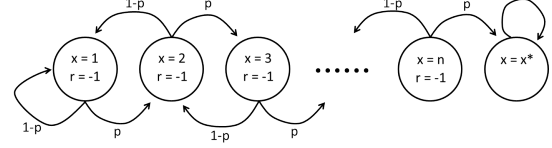


Figure 1. A Markov chain

6. Experiments

In this section we present numerical simulations of policy evaluation on a challenging continuous maze domain. The goal of this presentation is twofold; first, we show that the variance function may be estimated successfully on a large domain using a reasonable amount of samples. Second, the intuitive maze domain highlights the information that may be gleaned from the variance function. We begin by describing the domain and then present our policy evaluation results.

The Pinball Domain (Konidaris & Barto, 2009) is a continuous 2-dimensional maze where a small ball needs to be maneuvered between obstacles to reach some target area, as depicted in figure 3 (left). The ball is controlled by applying a constant force in one of the 4 directions at each time step, which causes acceleration in the respective direction. In addition, the ball's velocity is susceptible to additive Gaussian noise (zero mean, standard deviation 0.03) and friction (drag coefficient 0.995). The state of the ball is thus 4-dimensional (x, y, \dot{x}, \dot{y}) , and the action set is discrete, with 4 available controls. The obstacles are sharply shaped and fully elastic, and collisions cause the ball to bounce. As noted in (Konidaris & Barto, 2009), the sharp obstacles and continuous dynamics make the pinball domain more challenging for RL than simple navigation tasks or typical benchmarks like Acrobot.

A Java implementation of the pinball domain used in (Konidaris & Barto, 2009) is available on-line⁵ and was used for our simulations as well, with the addition of noise to the velocity.

We obtained a near-optimal policy using SARSA (Sutton & Barto, 1998) with radial basis function features and a reward of -1 for all states until reaching the target. The value function for this policy is plotted in

⁵<http://people.csail.mit.edu/gdk/software.html>

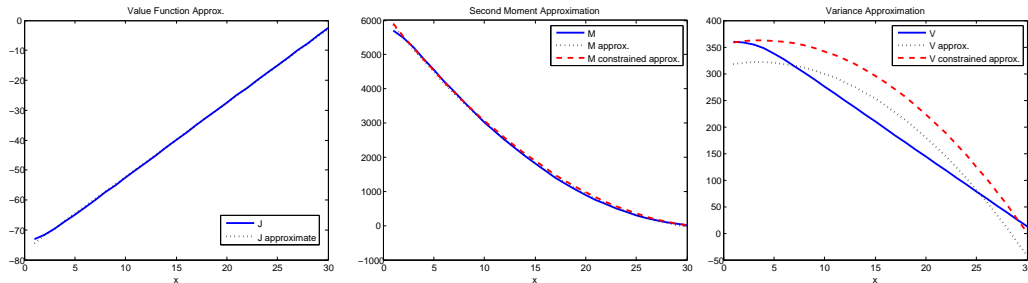


Figure 2. Value, second moment and variance approximation

Figure 3, for states with zero velocity. As should be expected, the value is approximately a linear function of the distance to the target.

Using 3000 trajectories (starting from uniformly distributed random states in the maze) we estimated the value and second moment functions by the LSTD(λ) algorithm described above. We used uniform tile coding as features (50×50 non-overlapping tiles in x and y , no dependence on velocity) and set $\lambda = 0.9$. The resulting estimated standard deviation function is shown in Figure 4 (left). In comparison, the standard deviation function shown in Figure 4 (right) was estimated by the naive sample variance, and required 500 trajectories from each point - a total of 1,250,000 trajectories.

Note that the variance function is clearly not a linear function of the distance to the target, and in some places not even monotone. Furthermore, we see that an area in the top part of the maze before the first turn is very risky, even more than the farthest point from the target. We stress that this information cannot be gleaned from inspecting the value function alone.

7. Conclusion

This work presented a novel framework for policy evaluation in RL with variance related performance criteria. We presented both formal guarantees and empirical evidence that this approach is useful in problems with a large state space.

A few issues are in need of further investigation. First, we note a possible extension to other risk measures such as the percentile criterion (Delage & Mannor, 2010). In a recent work, Morimura et al. (2012) derived Bellman equations for the *distribution* of the total return, and appropriate TD learning rules were proposed, albeit without function approximation and formal guarantees.

More importantly, at the moment it remains unclear how the variance function may be used for *policy optimization*. While a naive policy improvement step may be performed, its usefulness should be questioned, as it was shown to be problematic for the standard deviation adjusted reward (Sobel, 1982) and the variance constrained reward (Mannor & Tsitsiklis, 2011). In (Tamar et al., 2012), a policy gradient approach was proposed for handling variance related criteria, which may be extended to an actor-critic method by using the variance function presented here.

References

- Bertsekas, D. P. *Dynamic Programming and Optimal Control, Vol II*. Athena Scientific, fourth edition, 2012.
- Bertsekas, D. P. and Tsitsiklis, J. N. *Neuro-dynamic programming*. Athena Scientific, 1996.
- Bertsekas, D.P. Temporal difference methods for general projected equations. *IEEE Trans. Auto. Control*, 56(9):2128–2139, 2011.
- Borkar, V.S. *Stochastic approximation: a dynamical systems viewpoint*. Cambridge Univ Press, 2008.
- Boyan, J.A. Technical update: Least-squares temporal difference learning. *Machine Learning*, 49(2):233–246, 2002.
- Delage, E. and Mannor, S. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203–213, 2010.
- Horn, R. A. and Johnson, C. R. *Matrix Analysis*. Cambridge University Press, 1985.
- Howard, R. A. and Matheson, J. E. Risk-sensitive markov decision processes. *Management Science*, 18(7):356–369, 1972.

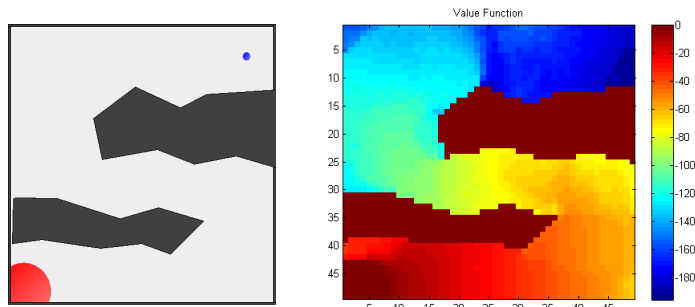


Figure 3. The pinball domain

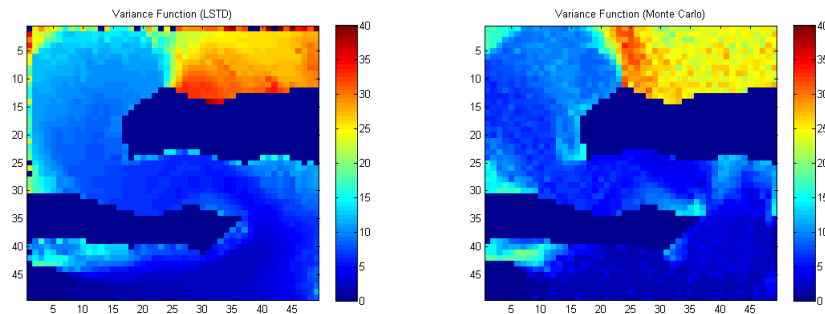


Figure 4. Standard Deviation of Reward To Go

Konidaris, G.D. and Barto, A.G. Skill discovery in continuous reinforcement learning domains using skill chaining. In *NIPS*, 2009.

Lazaric, A., Ghavamzadeh, M., and Munos, R. Finite-sample analysis of lstd. In *ICML*, 2010.

Luenberger, D. *Investment Science*. Oxford University Press, 1998.

Mannor, S. and Tsitsiklis, J. N. Mean-variance optimization in markov decision processes. In *ICML*, 2011.

Morimura, T., Sugiyama, M., Kashima, H., Hachiya, H., and Tanaka, T. Parametric return density estimation for reinforcement learning. *arXiv preprint arXiv:1203.3497*, 2012.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994.

Sharpe, W. F. Mutual fund performance. *The Journal of Business*, 39(1):119–138, 1966.

Sobel, M. J. The variance of discounted markov decision processes. *J. Applied Probability*, pp. 794–802, 1982.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning*. MIT Press, 1998.

Tamar, A., Di Castro, D., and Mannor, S. Policy gradients with variance related risk criteria. In *ICML*, 2012.

Tesauro, G. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

Supplementary Material

A. Proof of Proposition 2

Proof. The equation for $J(x)$ is well-known, and its proof is given here only for completeness. Choose $x \in X$. Then,

$$\begin{aligned}
 J(x) &= \mathbb{E}[B|x_0 = x] \\
 &= \mathbb{E}\left[\sum_{k=0}^{\tau-1} r(x_k) | x_0 = x\right] \\
 &= r(x) + \mathbb{E}\left[\sum_{k=1}^{\tau-1} r(x_k) | x_0 = x\right] \\
 &= r(x) + \mathbb{E}\left[\mathbb{E}\left[\sum_{k=1}^{\tau-1} r(x_k) | x_0 = x, x_1 = y\right]\right] \\
 &= r(x) + \sum_{y \in X} P(y|x)J(y)
 \end{aligned}$$

where we excluded the terminal state from the sum since reaching it ends the trajectory.

Similarly,

$$\begin{aligned}
 M(x) &= \mathbb{E}[B^2|x_0 = x] \\
 &= \mathbb{E}\left[\left(\sum_{k=0}^{\tau-1} r(x_k)\right)^2 | x_0 = x\right] \\
 &= \mathbb{E}\left[\left(r(x_0) + \sum_{k=1}^{\tau-1} r(x_k)\right)^2 | x_0 = x\right] \\
 &= r(x)^2 + 2r(x)\mathbb{E}\left[\sum_{k=1}^{\tau-1} r(x_k) | x_0 = x\right] + \mathbb{E}\left[\left(\sum_{k=1}^{\tau-1} r(x_k)\right)^2 | x_0 = x\right] \\
 &= r(x)^2 + 2r(x) \sum_{y \in X} P(y|x)J(y) + \sum_{y \in X} P(y|x)M(y).
 \end{aligned}$$

The uniqueness of the value function J for a proper policy is well known, c.f. proposition 3.2.1 in (Bertsekas, 2012). The uniqueness of M follows by observing that in the equation for M , M may be seen as the value function of an MDP with the same transitions but with reward $r(x)^2 + 2r(x) \sum_{y \in X} P(y|x)J(y)$. Since only the rewards change, the policy remains proper and proposition 3.2.1 in (Bertsekas, 2012) applies. \square

B. Proof of Lemma 8

Proof. We have

$$\begin{aligned}
 \|z_{true} - z^*\|_\alpha &\leq \|z_{true} - \Pi z_{true}\|_\alpha + \|\Pi z_{true} - z^*\|_\alpha \\
 &= \|z_{true} - \Pi z_{true}\|_\alpha + \|\Pi T z_{true} - \Pi T z^*\|_\alpha \\
 &\leq \|z_{true} - \Pi z_{true}\|_\alpha + \beta \|z_{true} - z^*\|_\alpha.
 \end{aligned}$$

rearranging gives the stated result. \square

C. Proof of Theorem 9

Proof. Let $\phi_1(x)$, $\phi_2(x)$ be some vector functions of the state. We claim that

$$\mathbb{E} \left[\sum_{t=0}^{\tau-1} \phi_1(x_t) \phi_2(x_t)^T \right] = \sum_x q(x) \phi_1(x) \phi_2(x)^T. \quad (17)$$

To see this, let $\mathbb{1}(\cdot)$ denote the indicator function and write

$$\begin{aligned} \mathbb{E} \left[\sum_{t=0}^{\tau-1} \phi_1(x_t) \phi_2(x_t)^T \right] &= \mathbb{E} \left[\sum_{t=0}^{\tau-1} \sum_x \phi_1(x) \phi_2(x)^T \mathbb{1}(x_t = x) \right] \\ &= \mathbb{E} \left[\sum_x \phi_1(x) \phi_2(x)^T \sum_{t=0}^{\tau-1} \mathbb{1}(x_t = x) \right] \\ &= \sum_x \phi_1(x) \phi_2(x)^T \mathbb{E} \left[\sum_{t=0}^{\tau-1} \mathbb{1}(x_t = x) \right]. \end{aligned}$$

Now, note that the last term on the right hand side is an expectation (over all possible trajectories) of the number of visits to a state x until reaching the terminal state, which is exactly $q(x)$ since

$$\begin{aligned} q(x) &= \sum_{t=0}^{\infty} P(x_t = x) \\ &= \sum_{t=0}^{\infty} \mathbb{E}[\mathbb{1}(x_t = x)] \\ &= \mathbb{E} \left[\sum_{t=0}^{\infty} \mathbb{1}(x_t = x) \right] \\ &= \mathbb{E} \left[\sum_{t=0}^{\tau-1} \mathbb{1}(x_t = x) \right], \end{aligned}$$

where the last equality follows from the absorbing property of the terminal state. Similarly, we have

$$\mathbb{E} \left[\sum_{t=0}^{\tau-1} \phi_1(x_t) \phi_2(x_{t+1})^T \right] = \sum_x \sum_y q(x) P(y|x) \phi_1(x) \phi_2(y)^T, \quad (18)$$

since

$$\begin{aligned} \mathbb{E} \left[\sum_{t=0}^{\tau-1} \phi_1(x_t) \phi_2(x_{t+1})^T \right] &= \mathbb{E} \left[\sum_{t=0}^{\tau-1} \sum_x \sum_y \phi_1(x) \phi_2(y)^T \mathbb{1}(x_t = x, x_{t+1} = y) \right] \\ &= \mathbb{E} \left[\sum_x \sum_y \phi_1(x) \phi_2(y)^T \sum_{t=0}^{\tau-1} \mathbb{1}(x_t = x, x_{t+1} = y) \right] \\ &= \sum_x \sum_y \phi_1(x) \phi_2(y)^T \mathbb{E} \left[\sum_{t=0}^{\tau-1} \mathbb{1}(x_t = x, x_{t+1} = y) \right] \end{aligned}$$

and

$$\begin{aligned}
 q(x)P(y|x) &= \sum_{t=0}^{\infty} P(x_t = x)P(y|x) \\
 &= \sum_{t=0}^{\infty} P(x_t = x, x_{t+1} = y) \\
 &= \sum_{t=0}^{\infty} \mathbb{E}[\mathbb{1}(x_t = x, x_{t+1} = y)] \\
 &= \mathbb{E} \left[\sum_{t=0}^{\infty} \mathbb{1}(x_t = x, x_{t+1} = y) \right] \\
 &= \mathbb{E} \left[\sum_{t=0}^{\tau-1} \mathbb{1}(x_t = x, x_{t+1} = y) \right].
 \end{aligned}$$

Since trajectories between visits to the recurrent state are statistically independent, the law of large numbers together with the expressions in (17) and (18) suggest that the approximate expressions in (12) converge to their expected values with probability 1, therefore we have

$$\begin{aligned}
 A_N &\rightarrow A, & b_N &\rightarrow b, \\
 C_N &\rightarrow C, & d_N &\rightarrow D,
 \end{aligned}$$

and

$$\begin{aligned}
 \hat{w}_{J;N}^* &= A_N^{-1} b_N \rightarrow A^{-1} b = w_J^*, \\
 \hat{w}_{M;N}^* &= C_N^{-1} d_N \rightarrow C^{-1} d = w_M^*.
 \end{aligned}$$

□

D. Proof of Theorem 10

Proof. Using (17) and (18) we have for all k

$$\begin{aligned}
 \mathbb{E} \left[\sum_{t=0}^{\tau^k-1} \phi_J(x_t) \delta_J^k(t, w_J, w_M) \right] &= \Phi_J^T Q r - \Phi_J^T Q (I - P) \Phi_J w_J, \\
 \mathbb{E} \left[\sum_{t=0}^{\tau^k-1} \phi_M(x_t) \delta_M^k(t, w_J, w_M) \right] &= \Phi_M^T Q R (r + 2P \Phi_J w_J) - \Phi_M^T Q (I - P) \Phi_M w_M,
 \end{aligned} \tag{19}$$

Letting $\hat{w}_k = (\hat{w}_{J;k}, \hat{w}_{M;k})$ denote a concatenated weight vector in the joint space $\mathbb{R}^{S_J} \times \mathbb{R}^{S_M}$ we can write the TD algorithm in a stochastic approximation form as

$$\hat{w}_{k+1} = \hat{w}_k + \xi_k (z + M \hat{w}_k + \delta M_{k+1}), \tag{20}$$

where

$$\begin{aligned}
 M &= \begin{pmatrix} \Phi_J^T Q (P - I) \Phi_J & 0 \\ 2\Phi_M^T Q R P \Phi_J & \Phi_M^T Q (P - I) \Phi_M \end{pmatrix}, \\
 z &= \begin{pmatrix} \Phi_J^T Q r \\ \Phi_M^T Q R r \end{pmatrix},
 \end{aligned}$$

and the noise terms δM_{k+1} satisfy

$$\mathbb{E} [\delta M_{k+1} | F_n] = 0,$$

where F_n is the filtration $F_n = \sigma(\hat{w}_m, \delta M_m, m \leq n)$, since different trajectories are independent.

We first claim that the eigenvalues of M have a negative real part. To see this, observe that M is block triangular, and its eigenvalues are just the eigenvalues of $\Phi_J^T Q (P - I) \Phi_J$ and $\Phi_M^T Q (P - I) \Phi_M$. By Lemma 6.10 in (Bertsekas & Tsitsiklis, 1996) these matrices are negative definite. It therefore follows (see Bertsekas, 2012 example 6.6) that their eigenvalues have a negative real part. Thus, the eigenvalues of M have a negative real part.

Next, let $h(w) = Mw + z$, and observe that the following conditions hold.

A 1. *The map h is Lipschitz.*

A 2. *The step sizes satisfy*

$$\sum_{k=0}^{\infty} \xi_k = \infty, \quad \sum_{k=0}^{\infty} \xi_k^2 < \infty.$$

A 3. *$\{\delta M_n\}$ is a martingale difference sequence, i.e., $\mathbb{E}[\delta M_{n+1} | F_n] = 0$.*

The next condition also holds

A 4. *The functions $h_c(w) \triangleq h(cw)/c$, $c \geq 1$ satisfy $h_c(w) \rightarrow h_\infty(w)$ as $c \rightarrow \infty$, uniformly on compacts, and $h_\infty(w)$ is continuous. Furthermore, the Ordinary Differential Equation (ODE)*

$$\dot{w}(t) = h_\infty(w(t))$$

has the origin as its unique globally asymptotically stable equilibrium.

This is easily verified by noting that $h(cw)/c = Mw + c^{-1}z$, and since z is finite, $h_c(w)$ converges uniformly as $c \rightarrow \infty$ to $h_\infty(w) = Mw$. The stability of the origin is guaranteed since the eigenvalues of M have a negative real part.

Theorem 7 in Chapter 3 of (Borkar, 2008) states that if A1 - A4 hold, the following condition holds

A 5. *The iterates of (20) remain bounded almost surely, i.e., $\sup_k \|\hat{w}_k\| < \infty$, a.s.*

Finally, we use a standard stochastic approximation result that, given that the above conditions hold, relates the convergence of the iterates of (20) with the asymptotic behavior of the ODE

$$\dot{w}(t) = h(w(t)). \tag{21}$$

Since the eigenvalues of M have a negative real part, (21) has a unique globally asymptotically stable equilibrium point, which by (10) is exactly $\hat{w}^* = (\hat{w}_J^*, \hat{w}_M^*)$. Formally, by Theorem 2 in Chapter 2 of (Borkar, 2008) we have that if A1 - A3 and A5 hold, then $\hat{w}_k \rightarrow \hat{w}^*$ as $k \rightarrow \infty$ with probability 1. \square